Serial No. 553,161

Filing Date 7 November 1995

Inventor George S. Kang
Lawrence J. Fransen.

NOTICE

The above identified patent application is available for licensing. Requests for information should be addressed to:

OFFICE OF NAVAL RESEARCH
DEPARTMENT OF THE NAVY
CODE OOCC3
ARLINGTON VA 22217-5660

Serial No.:
Inventors: Kang and Fransen

PATENT APPLICATION
Navy Case No. 77,023

METHOD AND APPARATUS FOR SEGMENTING A SPEECH WAVEFORM

BACKGROUND OF THE INVENTION

1. Field of the Invention

The present invention is directed to a system for processing human speech and, more particularly, to a system that pitch-synchronously segments the human speech waveform into individual pitch waveforms which may be transformed, replicated, and concatenated to generate continuous speech with desired speech characteristics.

2. Description of the Related Art

The ability to alter speech characteristics is important in both military and civilian applications with the increased use of synthesized speech in communication terminals, message devices, virtual-reality environments, and training aids. Currently, however, there is no known method capable of modifying utterance rate, pitch period, or resonant frequencies of speech by operating directly on the original speech waveform.

Typical speech analysis and synthesis are based on a model that includes a vocal tract component consisting of an electrical filter and a glottis component consisting of an excitation signal which is usually an electrical signal generator feeding the filter. A goal of these models is to convert the complex speech waveform into a set of perceptually significant parameters. By controlling these parameters, speech can be generated with these models. To derive human speech model parameters accurately, both the model input (turbulent air from the lungs) and the model output (speech waveform) are required. In conventional speech models, however, model parameters are derived using only the model output because the model input is not accessible. As a result, the estimated model parameters are not often accurate.

What is needed is a different way of representing speech that does not represent speech as an electrical analog sound production mechanism.

SUMMARY OF THE INVENTION

It is an object of the present invention to represent the speech waveform directly by individual waveforms beginning and ending with the pitch epoch. These waveforms will be referred to as pitch waveforms.

It is another object of the present invention to segment the speech waveform into pitch waveforms.

It is also an object of the present invention to perform pitch synchronous segmentation to obtain pitch waveforms by estimating the center of a pitch period by means of centroid analysis.

It is an additional object of the present invention to use the ability to segment the speech waveform to perform speech analysis/synthesis, speech disguise or change, articulation change, boosting or enhancement, timber change and pitch change.

It is an object of the present invention to utilize segmented pitch waveforms to perform speech encoding, speech recognition, speaker verification and text to speech.

It is a further object of the present invention to provide a speech model that is not affected by pitch interference, that is, segmented pitch waveform spectrum is free of pitch harmonics.

The above objects can be attained by a system that uses an estimate of the pitch period and an estimation of the center of the pitch waveform to segment the speech waveform into pitch waveforms. The center of the pitch waveform is determined by finding the centroid of the speech waveform for one pitch period. The centroid is found by finding a local minimum in the centroid histogram waveform, such that the local minimum corresponds to the midpoint of the pitch waveform. The midpoint or center of the pitch waveform along with the pitch period is used to segment or divide the speech waveform. The speech waveform can then be represented by a set of such pitch waveforms. The pitch waveform can be modified by frequency enhancement/filtering, waveform stretching/shrinking in speech synthesis. The utterance rate of the speech can also be changed by increasing or decreasing the number of pitch waveforms in the output.

BRIEF DESCRIPTION OF THE DRAWINGS

These and other objects, features and advantages of the invention, as well as the invention itself, will become better understood by reference to the following detailed description when considered in connection with the accompanying drawings wherein like reference numerals designate identical or corresponding parts throughout the several views and wherein:

Figure 1 depicts a speech waveform with delineated pitch waveforms and an associated pitch period;

Figures 2(a) and 2(b) respectively depict a low-pass filtered speech waveform and a centroid histogram waveform for the speech waveform;

Figure 3 depicts the typical hardware of the present invention in a preferred embodiment;

Figure 4 shows the pitch synchronous segmentation operation of the present invention performed by the computer system 40 of figure 3;

Figures 5(a), 5(b) and 5(c) illustrate utterance rate changes;

Figure 6 illustrates pitch waveform replication to generate continuous speech;

Figures 7(a), 7(b) and 7(c) depict pitch alteration;

Figure 8 depicts spectrum modifications;

Figure 9 shows the structural elements in the computer system 40 of figure 3 for performing the operations of segmenting and reconstructing a speech waveform;

Figure 10(a) illustrates timing circuits for generating various timing signals used in the system of figure 11;

Figure 10(b) depicts control timing diagrams;

Figure 11 depicts a discrete component embodiment of the invention;

Figure 12 depicts a first type of circuit for utilizing the segmented pitch waveform samples of figure 11 to modify the waveform spectrum;

Figure 13 depicts a second type of circuit for utilizing the segmented pitch waveform samples of figure 11 to modify the pitch;

Figure 14 depicts the components used for replicating and concatenating pitch waveforms to generate continuous analog speech;

Figure 15 illustrates an alternate approach to segmentation; and

Figure 16 depicts different functions used in correlation.

DESCRIPTION OF THE PREFERRED EMBODIMENTS

This invention is directed toward a speech analysis/synthesis model that characterizes the original speech waveform. In this invention, the speech waveform is modeled as a collection of disjoint waveforms, each representing a pitch waveform. Note that a pitch period is the time duration or the number of speech samples present in the pitch waveform. A segmented waveform of one pitch period can represent neighboring pitch waveforms because of the redundancy inherent in speech. Speech is reconstructed by replicating within a frame and concatenating from frame to frame the segmented pitch waveforms.

The present invention is a technique for segmenting pitch waveforms from a speech waveform of a person based on pitch. The inventors have recognized that the speech waveform, as illustrated in figure 1, is a collection of disjoint waveforms 1-10 where waveforms 2-10 are the result of the glottis opening and closing at the pitch rate. A purpose of the invention is to segment individual pitch waveforms. As noted previously, a segmented waveform of one pitch period T 123 can represent neighboring pitch waveforms because of the slowly varying nature of the speech waveform. One pitch waveform representing more than one pitch waveform is an important aspect of speech compression and synthesis. In the example of figure 1, eight pitch waveforms 3-10 are substantially similar. Because one pitch waveform represents speech that may have many pitch waveforms, speech compression is possible. Modification of the speech waveform can be accomplished by modifying only a portion of the speech waveform or the representative pitch waveform, an advantage in speech modification and synthesis.

The segmentation of a speech waveform into pitch waveforms requires two computational steps: 1) the determination of the pitch period; and 2) the determination of a starting point for the pitch waveform. Determining the pitch period can be performed using conventional techniques typically found in devices called vocoders. To determine the starting point of the pitch waveform, the center of the pitch waveform is determined first, in accordance with the present invention, by centroid histogram waveform analysis.

The location of the centroid (or center of gravity), as used in mechanics, is expressed by centroid function

$$\eta = \frac{\int_{x1}^{x2} x f(x)\, dx}{\int_{x1}^{x2} f(x)\, dx} \quad for \quad x1 \le x \le x2 \qquad (1)$$

where f(x) is a non-negative mass distribution, and [x1, x2] is the domain of variable x. In this invention, x is a time variable, f(x) is the speech waveform, and the domain [x1, x2] where $x_2 - x_1$, is one pitch period. Since f(x) cannot be negative, a sufficient amount of bias is added to the speech waveform so that f(x) > 0.

Figures 2(a) and 2(b) respectively show a low pass filtered speech samples S'(.) 126 and centroid histogram samples C(.) 182 of the centroid function. The center of a pitch period is defined to be at a local minimum of the centroid histogram samples C(.) 182 of the speech waveform for that pitch period. A local minimum location α 199 of the samples C(.) 182 occurs at a midpoint 20 of a pitch period T 123. Knowing the location α 199 and the pitch period T 123 allows a pitch period starting point β 24 and a pitch period ending point 26 of the pitch period T 123 to be determined. The pitch period starting point β 24 and ending point 26 define the boundaries of the pitch period T 123. By using the centroid to determine the segmentation, the present invention results in a balancing of the "weight" or left and right moments of the pitch waveform samples around the centroid.

The segmentation of the waveform samples, in accordance with the present invention, is preferably performed using a system 30 as illustrated in figure 3. An input analog speech signal 32, such as a human voice that is to be compressed or modified, from an input device 34, such as a microphone, is sampled by a conventional analog-to-digital converter (ADC) 36 at a conventional sampling rate suitable for speech processing. Digitized speech samples 38 are provided to a conventional computer system 40, such as a Sun workstation or a desktop personal computer. The computer system 40 performs the segmentation (as indicated in Fig. 4 - to be discussed) and any analysis or processing required for speaker verification, speaker recognition, text to speech, compression, modification, synthesis, etc. The segmented waveform in modified or unmodified form can be stored in a memory (disk or RAM - not shown) of the system 40. If the waveform is being modified, such as when disguised speech is to be produced, modified speech waveform 42 samples are converted by a conventional digital-to-analog converter (DAC) 44 into an analog speech signal 46 and provided to an output device 48, such as a speaker. The process of the present invention can also be stored in a portable medium, such as a disk, and carried from system to system.

The segmentation operation segments by determining the centroid, which is performed by the computer system 40, as illustrated in detail in figure 4. This segmentation operation starts by generating a ramp R(.) 171 of one pitch period duration in a ramp function generator 50. More specifically, the generator 50 is responsive to a pitch period T 123 for generating ramp R(.) 171, expressed by

$$R(i) = \begin{cases} i & \text{for } -T/2 \leq i \leq T/2 - 1 \\ 0 & \text{elsewhere} \end{cases} \qquad (2)$$

The ramp R(.) 171 is then correlated in a correlator 52 with low pass filtered speech samples S'(.) 126 to produce a centroid function or the centroid histogram samples C(.) 182. The use of low pass filtered speech samples S'(.) 126 is preferred because it is free of high frequency information often present in the speech waveform. By definition, a centroid function is the sum of the products of the ramp R(.) 171 samples and the low-pass filtered speech samples S'(.) 126 with a successive mutual delay (which is a cross correlation function). Thus, the centroid histogram samples C(.) 182 are expressed by

$$C(i) = \sum_{j=-T/2}^{T/2-1} R(j)S'(i+j) \qquad L - T/2 \le i \le L + T/2 - 1$$

$$= \sum_{j=-T/2}^{T/2-1} jS'(i+j) \qquad L - T/2 \le i \le L + T/2 - 1 \qquad (3)$$

where L is the midpoint of the centroid analysis frame. As noted from the above expression (3), the samples C(.) 182 are computed for one pitch period around the center of the analysis frame. The typical centroid histogram samples C(.) 182 waveform is illustrated in figure 2(b), which was previously discussed.

Next, a local minimum search 54 is performed on the samples C(.) 182 (see figure 4) to determine a local minimum location α. As previously noted, the midpoint of the pitch waveform coincides with the local minimum of the samples C(.) 182. This minimum location, denoted by α 199, is obtained from

$$\alpha = \min_{i} \{C(i)\} \qquad L - T/2 \le i \le L + T/2 - 1 \tag{4}$$

As illustrated in figures 2(a) and 2(b), the minimum location α 199 corresponds to the midpoint 20 of the pitch period T 123. Thus, the pitch epoch begins at $\alpha - T/2$, the pitch period starting point β 24, and ends at $\alpha + T/2 - 1$ or pitch period ending point 26.

The minimum location α 199 needs refinement because the pitch period T 123 provided by a pitch tracker 121 (figure 9) is often not too accurate. Thus, both the local minimum location α 199 and pitch period T 123 are refined by repeating the above local minimum search 54 for each $T \pm \Delta T$ where ΔT is as much as T/16 (6.25% of T). This refinement 56 improves the segmentation performance. The refined local minimum location and refined pitch period are denoted by α' and T', respectively.

Finally, segmented pitch waveform samples 131 are excised from the speech samples S(.) 128 by a switch 58 from time $\alpha' - T'/2$ to time $\alpha' + T'/2 - 1$.

Once the segmented pitch waveform samples 131 are excised they can be modified, replicated, etc., as will be discussed in more detail later, to produce a reconstructed speech waveform. This is accomplished by replicating and concatenating pitch waveforms. Because the synthesis frame size M is generally greater than the pitch period T', the segmented speech waveform is usually replicated more than once. The segmented waveform is always replicated in its entirety. Near the boundary of the synthesis frame, it is necessary to decide whether the segmented waveform of the current frame should be replicated again or the segmented waveform of the next frame should be copied. The choice is determined by the remaining space in relation to the length of the segmented waveform T'. If the remaining space is greater than T'/2, the segmented waveform of the current frame is replicated again. On the other hand, if the remaining space is less than or equal to T'/2, the segmented waveform of the next frame is copied. Any significant discontinuity at either of the segmented pitch waveform boundarys 24 and 26 (figure 2(a)) will produce clicks or warbles in the reconstructed speech. To avoid discontinuities, the system performs a three-point interpolation at the pitch epoch (see FORTRAN program on pages A1 through A10 of the Appendix for details of this operation).

As noted previously the segmented pitch waveform samples 131 can be used for speaker verification, speaker recognition, text to speech, compression, synthesis or modification of the speech waveform. The modification operation can independently modify the utterance rate, the pitch, and the resonance frequencies of the original speech waveform.

Referring now to figures 5(a) - 5(c), the speech utterance rate is altered by simply changing the number of pitch waveforms replicated at the output. Therefore, the utterance rate is controlled by synthesis frame size, M, relative to the analysis frame size, N, which are internal parameters that can be altered by the operator. The relationship of N and M are shown in figure 6. Three cases for the relationship between N and M are: (1) M = N: In this case, the utterance rate is unaltered because the same number of pitch waveforms are present in both the input and output frames (see figure 5(a)). (2) M > N: The output speech is slowed down by replicating the pitch waveform 60 more than once, producing replicated waveforms 62 used to fill the synthesis frame M (see figure 5(b) for an example). (3) M < N: The output speech is sped up because the output frame M has fewer pitch waveforms than the input frame N (see figure 5(c)). In these examples of utterance rate change the pitch period and resonance frequencies of the original speech are not affected by modifying the speech rate.

Pitch can be changed by expanding or compressing the pitch waveform. Alteration of the pitch period in this invention does not affect the speech utterance rate, but resonant frequencies do change in proportion to the pitch. It is common knowledge that high-pitch female voices have higher resonant frequencies than low-pitch male voices for the same vowel. The natural coupling of pitch frequency and resonant frequencies is beneficial.

Figures 7(a)-7(c) illustrate the effect of changing the pitch. Figure 7(a) shows the original speech. Figure 7(b) is altered speech played back with a 30% lower pitch. Figure 7(c) is altered speech played back with a 30% higher pitch.

The resonant frequencies of speech can be modified by altering the pitch waveform spectrum. An example of such an alteration is illustrated in figure 8. In step 1, a conventional discrete Fourier transform (DFT) is applied to the segmented pitch waveform samples 131 to produce an amplitude spectrum 74. In step 2, the spectrum 74 is modified in some conventional manner to produce a modified amplitude spectrum 78. For example, the first resonant frequency in the spectrum 74 can be shifted to the left as shown by the spectrum 78. In step 3, a conventional Hilbert transformation is performed on spectrum 78 to produce a modified phase spectrum 82. In step 4, an inverse discrete Fourier transform (IDFT) is performed on amplitude spectrum 78 and phase spectrum 82 to produce a modified pitch waveform 135 with altered spectral characteristics. This waveform 135 can then be used to generate speech.

Figure 9 shows the structural elements in the computer system 40 of figure 3 for performing the operations of segmenting and reconstructing a speech waveform. As shown in figure 9, three inputs 111, 113 and 32 are provided: an analysis frame size N 111 (an integer from 60 to 240), a synthesis frame size M 113 (an integer from 60 to 240) and the input analog speech signal 32. The analysis frame size N 111 and the synthesis frame size M 113 are provided by the operator prior to start up of system 30 (figure 3). The analog signal 32 from an input device, such as a microphone, is converted by the ADC 36 into a series of digitized speech samples 38 supplied at an 8-kHz rate. Although not shown, the analog signal 32 is low pass filtered by the ADC 36 prior to the conversion to pass only signals below 4 kHz. The digitized speech samples 38 are conventionally filtered by a low-pass filter 119 to pass low pass filtered speech samples 120 at audio frequencies below about 1 kHz while the original signal is delayed in a shift register 125 (to be discussed) to produce delayed speech samples S(.) 128. The pitch of the low pass filtered speech samples 120 is pitch period tracked by a conventional pitch tracker 121 to produce the pitch period T 123 (figure 4). A conventional pitch tracker is described in Digital Processing Of Speech Signals, by Rabiner et al, Prentice-Hall, Inc., NJ 1978, Chapter 4. The low pass filtered speech samples 120 are delayed in a shift register 127 (to be discussed). The delays of shift registers 125 and 127 are preselected to time align the low pass filtered speech samples S'(.) 126 and speech samples S(.) 128 with the pitch period T 123 for input to a pitch-synchronous speech segmentor 129. The pitch period T 123, the low pass filtered speech samples S'(.) 126, the speech samples S(.) 128 and the analysis frame size N 111 are used to perform segmentation in the segmentor 129 of the original signal as described with respect to figure 4. The segmented pitch waveform samples 131 are then transformed in an application dependent pitch waveform transformator 133, in one or more of the ways as previously discussed, to produce the modified pitch waveform 135. Speech is reconstructed in a speech waveform reconstructor 139 using the modified pitch waveform 135 and the synthesis frame size M 113 to produce the modified speech waveform 42. The modified speech waveform 42 is converted by DAC 44 into the output analog speech signal 46 which is supplied to an output device, such as a speaker (not shown).

The operations of figure 9, including the segmentation of figure 4, are described in more detail in the FORTRAN source code Appendix included herein.

In a hardware embodiment of the invention, the pitch synchronous segmentation of speech in the present invention can also be performed by an exemplary system 148 using discrete hardware components, as illustrated in figure 11. In the exemplary system 148, the minimum location and pitch period refinement 56 (figure 4) is not performed. Also, the analysis frame size N 111 is restricted to the range where $160 \le N \le 240$.

Before figure 11 is discussed, reference will now be made to figures 10(a) and 10(b). Figure 10(a) illustrates timing circuits for generating the various timing signals used in the system of figure 11, and figure 10(b) illustrates the control timing signals generated by the timing circuits of figure 10(a).

In figure 10(a), a clock generator 136 generates eight mega Hertz (8 MHz) clocks which are applied to an upper input of an AND gate 138 and to a 1000:1 frequency count down circuit 140. At this time the AND gate 138 is disabled by a 0 state signal from the Q output of a flip flop 142. The 8 MHz clocks are continuously counted down by the 1000:1 frequency count down circuit 140 to generate an 8 kHz speech sampling clock A (shown in figure 10(b)) each time that the count down circuit 140 counts 1000 8 MHz clocks and then is internally reset to zero (0) by the 1000 th 8 MHz clock. Note that the interpulse period of clock A is 125 microseconds (μs).

The 8 kHz speech sampling clock A is applied to M:1 and N:1 frequency count down circuits 144 and 146. It will be recalled that the synthesis frame size M 113 and the analysis frame size N 111 are internal parameters that can be altered by the operator. Thus, the values of M and N are selected by the operator.

The 8 kHz clock A is counted down by the M:1 frequency count down circuit 144 to generate an 8 kHz/M synthesis frame clock C (shown in figure 10 (b)) each time that the count down circuit 144 counts M 8 kHz A clocks and then is internally reset to 0 by the M th 8 kHz clock. In a similar manner, the 8 kHz clock A is counted down by the N:1 frequency count down circuit 146 to generate an 8 kHz/N analysis frame clock B (shown in figure 10(b)) each time that the count down circuit 146 counts N 8 kHz A clocks and then is internally reset to 0 by the N th 8kHz clock.

The 8 kHz/N analysis frame clock B is also applied to a 25 μs delay circuit 147 to produce a selected centroid histogram samples transfer signal E (shown in figure 10 (b)) which occurs 25 μs after each B clock. In a similar manner, the 8 kHz/N B clock is applied to a 50 μs delay circuit 150 to produce a begin pitch waveform modification signal F (shown in figure 10 (b)) which occurs 50 μs after the B clock. The B clock is also applied to a 100 μs delay circuit 152 to produce a ramp transfer signal D (shown in figure 10 (b)) which occurs 100 μs after the B clock.

Each time that an F clock is generated by the 50 μs delay circuit 150, that F clock sets the flip flop 142 to cause the Q output of the flip flop 142 to change to a 1 state output. This 1 state output enables the AND gate 138 to pass 8 MHz clocks. These 8 MHz clocks from AND gate 138 will henceforth be called T_2 pulses G, which will be applied to a shift register 195 in figure 11 (to be discussed).

The T_2 pulses G from AND gate 138 are counted by a T_2:1 frequency count down circuit 154 to generate a T_2:1 signal each time that the count down circuit 154 counts T_2 pulses and then is internally reset to 0 by the T_2 th 8 MHz clock that occurs after the flip flop 142 is set. The T_2:1 clock also resets the flip flop 142 so that the Q output of the flip flop 142 changes to a 0 state to disable the AND gate 138. Thus, no more T_2 pulses G are supplied to the shift register 195 in figure 11 at this time.

As shown in figure 10(b), the T_2 8 MHz pulses G start with the generation of the begin pitch waveform modification signal F and terminate after the frequency count down circuit 154 has counted T_2 8 MHz pulses G after the generation of the F signal.

Referring back to figure 11, the parameter analysis frame size N 111 signal is applied to shift registers 183 and 191, switches 187 and 193, minimum locator 189, and parallel-to-serial shift register 195. Speech samples S(.) 128 are fed at the time of the A clocks through AND gate 155 to the shift register 191. Pitch period T 123 signal is fed at the time of the D clocks through AND gate 159 to a shift register 167 and a ramp generator 169. The low-pass filtered (LPF) speech samples S'(.) 126 are fed at the time of the A clock through AND gate 163 to a shift register 165.

A conventional pitch tracker 121 (figure 9) used for this embodiment is able to track pitch with a range of 51 Hz to 400 Hz. A low-pitch male voice of 51 Hz corresponds to a pitch period, T, of 156 speech samples. A high-pitch female voice of 400 Hz corresponds to a pitch period, T, of 20 speech samples. Thus, the segmentation process must be able to handle pitch waveforms having 20 to 156 speech samples. Shift register 165 retains 156 filtered speech samples S'(.) 126, and shift register 175 stores ramp samples R_1 to R_T.

A ramp generator 169 develops an appropriate ramp R(.) 171 to be fed at the time of the ramp transfer signal D through an AND gate 173 to the shift register 175. The number of ramp samples transferred is T, and the appropriate ramp R_1 to R_T from the following list is transferred.

```
R1 to R20:  -10,-9,-8,...,-1,0,1,...,7,8,9
R1 to R21:  -10,-9,-8,...,-1,0,1,...,8,9,10
R1 to R22:  -11,-10,-9,...,-1,0,1,...,8,9,10
  ...          ...
  ...          ...
  ...          ...
R1 to R154:  -77,-76,-75,...,-1,0,1,...,74,75,76
R1 to R155:  -77,-76,-75,...,-1,0,1,...,75,76,77
R1 to R156:  -78,-77,-76,...,-1,0,1,...,75,76,77
```

Corresponding ramp samples R_1 to R_T from shift register 175 and corresponding low pass filter speech samples S'_1 to S'_T from shift register 165 are respectively cross multiplied in associated multipliers 166 to develop and apply cross products 179 to a summation unit 181.

Cross-products 179 of filtered speech samples S'_1 to S'_T with ramp R_1 to R_T pass through the summation unit 181 to form centroid histogram samples C(.) 182 for feeding into buffer 183. Ramp R_1 to R_T remains fixed over an analysis frame of N speech samples S(.) 128. An analysis frame of N filtered speech samples S'(.) 126 produces a frame of N sums of cross products designated

C_1 to C_N in register 183. C_1 to C_N is also designated frame 1 in register 183. Because a pitch waveform can spread over three frames, selection of a pitch waveform progresses from the middle of the three frames of register 183, at location 3N/2. Location 3N/2 is positioned in the middle of frame 2 of register 183. Because the search is now centered in frame 2, a one frame delay is introduced in the segmentation process. Register 167 delays the pitch one frame to properly line up the pitch in time with frame 2 of register 183.

Analysis frame size N 111 samples is fixed prior to the start up of ADC 36 and DAC 44 (ADC and DAC are shown in figures 3 and 9). Since N can range in the exemplary system 148 from 160 to 240 samples and the pitch period T 123 can range from 20 to 156 samples, three frames, 3N, of centroid samples are preserved in register 183.

The goal is to find a pitch waveform to associate with frame 2 of register 183. The beginning of the pitch cycle must be found such that a replication and concatenation process to be performed later will not create audible discontinuities. Each sum of cross products C_1 to C_N from summation unit 181 that is fed into register 183 is an indication of the center of gravity of a pitch waveform. The midpoint of a new pitch waveform occurs when the center of gravity is at a relative minimum.

A search window for locating the segmented pitch waveform samples 131 (of Figs. 4 and 11) is centered about the middle of frame 2. A search controller 197, such as a microprocessor, computes $\Delta=T_2/2$. The range of the search window is from centroid histogram sample $C_{3N/2-\Delta}$ to $C_{3N/2+\Delta}$ which encompasses $2\Delta+1$ samples, or a little greater than a pitch period of samples.

Once per analysis frame, centroid samples $C_{3N/2-\Delta}, \ldots, C_{3N/2+\Delta}$ are fed at the time of the E signal through AND gates 185 and through switch 187 to the minimum locator 189. Locator 189 is a conventional device, such as a microprocessor, used for finding the location of the minimum value of the centroid samples $C_{3N/2-\Delta}, \ldots, C_{3N/2+\Delta}$ within the locator 189. The pitch period starting point β 24 of the selected pitch waveform is in the range of $3N/2-\Delta$ to $3N/2+\Delta$. The starting point β 24 is passed to the switch 193. Switch 193 transfers T_2 speech samples from shift register 191 to shift register 195. Segmented pitch waveform samples 131 are available for the application dependent pitch waveform transformator 133. Shift register 191 has a size of 6N to have sufficient speech samples available.

Figure 12 shows a first type of circuit for utilizing the segmented pitch waveform samples 131 output of figure 11 to modify the waveform spectrum. In this circuit of figure 12, resonant frequencies of the segmented pitch waveform 131 are altered. The application of timing signal G (figures 10(a) and 10(b)) to the shift register 195 (figure 11) enables segmented pitch waveform samples 131 to be fed from the shift register 195 to a DFT unit 205. Amplitude and phase spectrum output 207 from DFT unit 205 are changed by an amplitude and phase spectrum modification unit 209 in a manner similar to that previously described in figure 8.

To explain this amplitude and phase spectrum modification being performed by circuit 209 of figure 12 reference will now be made back to the description of figure 8.

The resonant frequencies of speech can be modified by altering the pitch waveform spectrum. An example of altering the first resonant frequency is illustrated in figure 8. In step 1, a conventional DFT is applied to the segmented pitch waveform samples 131 to produce the amplitude spectrum 74. In step 2, the spectrum 74 is modified in some conventional manner to produce the modified amplitude spectrum 78. For example, the first resonant frequency in the spectrum 74 can be shifted to the left as shown by the spectrum 78. In step 3, a conventional Hilbert transformation is performed on spectrum 78 to produce the modified phase spectrum 82. In step 4, an IDFT is performed on amplitude spectrum 78 and phase spectrum 82 to produce the modified pitch waveform 135 with altered spectral characteristics. This waveform 135 can then be used to generate speech. This would tend to disguise speaker identity.

Now referring back to figure 12, a modified amplitude spectrum and phase spectrum signal 210 from the amplitude and phase spectrum modification unit 209 is inverted using an IDFT unit 211 and the resultant modified pitch waveform 135 is output to a 156 sample serial-to-parallel shift register 213.

Figure 12 can be changed to pass the segmented pitch waveform samples 131 unaltered through the circuit of figure 12 by removing the amplitude and phase spectrum modification circuit 209 and applying the output of the DFT unit 205 directly to the input of the IDFT unit 211 or by applying the output from shift register 195 (figure 11) directly to the input of shift register 213.

Another alternate embodiment of the discrete component version of this invention is illustrated in figure 13. The segmented pitch waveform samples 131 stored in shift register 195 pass through a stretching or shrinking transformation. Pitch waveform samples 131 are applied to a DAC 321 with the 8 kHz clock A (clock A generation is shown in figures 10(a) and 10(b)). The analog pitch waveform is resampled by an ADC 323 at a new sampling rate denoted by H (permissible values for H are 4 kHz $\leq$ H $\leq$ 16kHz) to create the modified pitch waveform 135 with T" samples stored in the shift register 213. Shrinking the pitch waveform raises the pitch, and expanding the pitch waveform lowers the pitch.

A discrete component waveform reconstruction circuit is illustrated in figure 14. This circuit comprises the shift register 213, a 156-sample, serial-to-parallel shift register 433, and two 156-sample, parallel-to-serial shift registers 431 and 435. Since the pitch period T 123 has a range of 20 to 156 samples, each of the 156-sample registers 213, 431, 433, and 435 enable the storage of the maximum number of samples in a pitch waveform.

A control circuit 445 generates 312 - T_2 pulses at an 8 MHz rate beginning at the time that clock E is generated. The control circuit 445 includes a flip flop 441 which is enabled by clock E to allow 8 MHz pulses to pass through an AND gate 437. A frequency count down circuit 439 permits 312-T_2 8 MHz pulses to pass through the AND gate 437 before it counts to a count of 312-T_2. When the frequency count down circuit 439 reaches a count of 312-T_2, it resets the flip flop 441 and internally resets itself to a 0 count. When reset, the Q output of the flip flop 441 changes to a 0 state to disable the AND gate 437. At this time no further 8 MHz pulses can be output from the control circuit 445 until the flip flop 441 is reset by the next enabling E clock.

Modified pitch waveform 135 samples are updated once per analysis frame. For purposes of this description, the updating operation of figure 14 will be described in relation to the utilization circuit of figure 12. However, it should be understood that a similar description of figure 14 is also applicable to the utilization circuit of figure 13.

In operation, modified pitch waveform 135 samples from figure 12 are serially clocked into serial-to-parallel register 213 by the G clock (figure 10(b)), which G clock is comprised of T_2 8 MHz clocks. At the time of the B clock, the stored samples in register 213 are shifted into and stored in parallel in the parallel-to-serial shift register 431. Since T_2 is often less than the 156-sample register-capacity of each of the registers 213 and 431, there is null data (i.e., data not related to the pitch waveform) comprising 156-T_2 samples positioned in time prior to the pitch waveform in the registers 213 and 431.

At the time of the next E clock, following the G clock during which the modified pitch waveform 135 samples were stored in the register 213, the flip flop 441 is set to enable AND gate 437 to pass 8 MHz clocks to registers 431 and 433. These 8 MHz clocks from AND gate 437 enable the samples stored in the register 431 to be serially clocked out of the register 431 into register 433. This transfer repositions the null data in time behind the speech data in register 433. More specifically, the first 156 clock pulses from the AND gate 437 in the circuit 445 transfer the entire contents of the register 431 to register 433, and the additional 156-T_2 clock pulses eliminate null data prior to the speech data in register 433.

The 8 MHz clocks from the AND gate 437 are also counted by a frequency count down circuit 439. When the circuit 439 reaches a count of ($312 - T_2$) 8 MHz clocks, it generates a signal to reset the flip flop 441 to disable the AND gate 437 so that no further 8 MHz clock pulses are output from the control circuit 445 until the flip flop 441 is set by the next enabling clock E.

The 8 kHz clock A is fed to a frequency count down circuit 443 to transfer in parallel the contents of register 433 to register 435 and to internally reset the counter 443 to zero (0) when the counter 443 has counted T_2 A clocks. Finally, T_2 samples of register 435 are fed at an 8 kHz rate by clock A to form the waveform 42 which is then applied to the DAC 44 at the A clock rate. The entire pitch waveform comprised of T_2 samples must be transferred in its entirety. The resulting analog speech signal 46 is then applied to the output device 48.

Additional details of uses for the present invention can be found in Naval Research Laboratory report NRL/FR/5550-94-9743 entitled Speech Analysis and Synthesis Based on Pitch-Synchronous Segmentation of the Speech Waveform by the inventors Kang and Fransen, published November 9, 1994 and available from Naval Research Laboratory, Washington, D.C. 20375-5320 and incorporated by reference herein.

The present invention is described with respect to performing pitch synchronous segmentation using centroid analysis, however, the segmentation can be performed in other ways. A direct approach is a method that determines pitch epochs directly from the waveform. An example of such an approach is peak picking in which the peaks 500 of the pitch waveforms are used to find the segment speech waveform. For certain speech waveforms, such an approach is feasible because the speech waveform shows pitch epochs rather clearly as in figure 15. One should be warned, however, that many speech waveforms do not show pitch epochs clearly. This is particularly true with non-resonant high-pitch female voices. As a result, this approach is not preferred.

Contrary to the direct method which uses instantaneous values of speech samples, a correlation method makes pitch epoch determination based on the ensemble averaging of a certain function derived from the speech waveform. The centroid method presented previously is a correlation process. The concept of the centroid originated in mechanical engineering to determine the center of gravity of a flat object. The concept of the center of gravity has been used in the field of signal analysis in recent years (See Papoulis A, Signal Analysis, McGraw-Hill Book Company, New York, New York 10017). For the speech waveform, the quantity x is a time variable, f(x) is the speech waveform, x1 is the pitch epoch, and x2-x1 is the current pitch period which is known beforehand. As elaborated in NRL Report

9743 (previously referenced), the above expression produces virtually identical pitch epoch locations as the following simplified expression:

$$\eta = \int_{x1}^{x2} x\, f(x)\, dx \qquad \text{for} \quad x1 \le x \le x2. \tag{6}$$

Thus, the centroid function is a cross correlation function between a ramp function and f(x). Ramp R(.) 171, as illustrated in figure 16, appearing in the above equation is odd-symmetric with respect to its midpoint. Other odd symmetric functions, such as a sine function 512 and a step function 514 of figure 16, can be used as a substitute for the ramp function. However, these alternative functions do not work as well as the ramp function and are thus not preferred.

The advantages of the present invention include the following. Speech utterance rate can be changed without altering the pitch or resonant frequencies. Pitch can be changed without altering the utterance rate. Resonant frequencies can be changed by spectrally shaping the pitch waveform without altering the utterance rate or pitch. The modified speech is similar to the original speech (not synthetic speech). Thus, the transformed speech intelligibility and quality are excellent. This invention has the feature of segmenting the speech waveform in terms of the pitch waveform. In the invention, the pitch waveform is a minimum inseparable entity of the speech waveform. Modification of the pitch waveform leads to speech characteristic alteration.

The many features and advantages of the invention are apparent from the detailed specification and, thus, it is intended to cover all such features and advantages of the invention which fall within the true spirit and scope of the invention. Further, since numerous modifications and changes will readily occur to those skilled in the art, it is not desired to limit the invention to the exact construction and operation illustrated and described, and accordingly all suitable modifications and equivalents may be resorted to, falling within the scope of the invention.

Serial No.:
Inventors: Kang and Fransen

PATENT APPLICATION
Navy Case No. 77,023

ABSTRACT OF THE DISCLOSURE

A system that synchronously segments a speech waveform using pitch period and a center of the pitch waveform. The pitch waveform center is determined by finding a local minimum of a centroid histogram waveform of the low-pass filtered speech waveform for one pitch period. The speech waveform can then be represented by one or more of such pitch waveforms or segments during speech compression, reconstruction or synthesis. The pitch waveform can be modified by frequency enhancement/filtering, waveform stretching/shrinking in speech synthesis or speech disguise. The utterance rate can also be controlled to speed up or slow down the speech.

Navy Case No. 77,023

Appendix
(FORTRAN Source Code)

```
c        NOTE  This program segments the speech waveform pitch
c              synchronously.  The segmented pitch waveform is
c              replicated and concatenated to generate
c              continuous speech.  The analysis frame size N
c              and synthesis frame size M are user specified.
c              Speech can be sped up by making N>M, or speech
c              may be slowed down by making N<M.
c
         integer T,Tprime,ubnd
         integer*2 is(240),idc(240),ilpf(240),ix(1)
         integer*2 i5dc(1200),i5lpf(1200)
         dimension amp(80),amp1(80),amp2(80),ampi(80)
         dimension phase(80),phase1(80),phase2(80),phasei(80)
         dimension pps(160),pw(160),xx(160)
         character*75 fname
c
c        ----- voice i/o setup -----
c
         write(6,1000)
1000     format('enter input speech file'/)
         read(5,1001) fname
1001     format(a)
c
c        *** initialize input device (not shown)
c
         write(6,1002)
1002     format('enter output speech file'/)
         read(5,1001) fname
c
c        *** initialize output device (not shown)
c
c
c        ----- initialization -----
c
c        *** analysis frame size N:   60<=N<=240
c
         N=100
c
c        *** synthesis frame size M:  60<=M<=240
c
         M=100
```

```
c
c       *** constants
c
        lpfoffset=9
        twopi=2.*3.14159
c
c       ----- input speech samples --------
c
c       *** transfer N speech samples into array is(.)
c       *** in indicates how many samples actually transferred
c       *** subroutine spchin is not shown
c
100     call spchin(N,is,in)
        if(in.eq.0) go to 999
        ifrmct=ifrmct+1
c
c       =========== preprocessing ==============
c
c       ----- remove dc from speech -----
c
        do 110 i=1,N
        x=is(i)
        call dcremove(x,y)
110     idc(i)=y
c
c       ----- store 5 dc-removed frames -----
c
        do 120 i=1,N
        i5dc(i)=i5dc(i+N)
        i5dc(i+N)=i5dc(i+2*N)
        i5dc(i+2*N)=i5dc(i+3*N)
        i5dc(i+3*N)=i5dc(i+4*N)
120     i5dc(i+4*N)=idc(i)
c
c       ----- low-pass filter -----
c
        do 130 i=1,N
        x=idc(i)
        call lpf(x,y)
130     ilpf(i)=y
```

```
c
c       ----- store 5 low-passed frames -----
c
        do 140 i=1,N
        i5lpf(i)=i5lpf(i+N)
        i5lpf(i+N)=i5lpf(i+2*N)
        i5lpf(i+2*N)=i5lpf(i+3*N)
        i5lpf(i+3*N)=i5lpf(i+4*N)
140     i5lpf(i+4*N)=ilpf(i)
c
c       =============== analysis ===============
c
c       ----- pitch tracker -----
c
c       NOTE    Use any reliable pitch tracker with an internal
c               two frame delay (pitch tracker not shown)
c
        call pitch(N,i5lpf,T)
        if(T.gt.128) T=128
c
c       ----- upper and lower bounds of search window -----
c
        icenter=2.5*N
        if(icenter.lt.T) icenter=T
        lbnd=icenter-.5*(T+1)
        ubnd=icenter+.5*(T+1)
c
c       ----- find pitch epoch and refine -----
c
        call centroid(lbnd,ubnd,T,i5lpf,small,loc)
        call adjust(T,loc,i5lpf,small,sadj,locadj,Tprime)
c
c       ----- compensate for lpf delay -----
c
        locadj=locadj-lpfoffset
c
c       ----- extract one pitch-waveform and compute rms -----
c
        index=locadj-Tprime/2
        if(index.ge.1) go to 150
        index=1
150     k=0
        sum=0.
        do 160 i=index,index+Tprime-1
        k=k+1
        pps(k)=i5dc(i)
160     sum=sum+pps(k)**2
        rms=sqrt(sum/Tprime)
```

```
c
c       NOTE   Introduce pitch modification here (expand or
c              compress pps(.) and change Tprime accordingly)
c
c       ----- Fourier transform the extracted pitch waveform-----
c
c       NOTE   The pitch waveform is interpolated in the
c              frequency domain during the intra-pitch period
c
        call dft(Tprime,pps,amp,phase,nn)
c
c       NOTE   Introduce spectrum modification here
c
        do 170 i=nn+1,80
        amp(i)=0.
170     phase(i)=0.
c
c       ----- store two frames of data -----
c
c       *** amplitude spectrum of pitch waveform
c
        do 180 i=1,80
        amp2(i)=amp1(i)
180     amp1(i)=amp(i)
c
c       *** phase spectrum of pitch waveform
c
        do 181 i=1,80
        phase2(i)=phase1(i)
181     phase1(i)=phase(i)
c
c       *** pitch period
c
        ipt2=ipt1
        ipt1=Tprime
c
c       *** pitch waveform rms
c
        irms2=irms1
        irms1=rms
```

```
c
c       ----- interpolation rate -----
c
c       NOTE    Use a faster interpolation if rms changes
c               significantly across frame boundary
c
        ratio=iabs(irms1-irms2)
        if(ratio.le.3.) ur=1.
        if(ratio.gt.3.and.ratio.le.6) ur=1.2
        if(ratio.gt.6) ur=1.4
c
c       =============== synthesizer ===============
c
        do 300 l=1,M
c
        if(im-ipti)240,200,200
200     im=0
c       ----- pitch epoch -----
c
c       NOTE    At each pitch epoch, amplitude normalize
c               the pitch waveform of the previous pitch
c               period and dump out sample by sample.
c
c       *** amplitude normalization factor
c
        sum=0.
        do 210 i=1,ipti
210     sum=sum+xx(i)**2
        gain=rmsi/sqrt(sum/ipti)
c
c       *** amplitude normalize past pitch waveform
c
        do 220 i=1,ipti
        u3=u2
        u2=u1
        u1=gain*xx(i)
c
c       *** perform 3-point interpolation only at pitch epoch
c
        u0=u2
        if(i.eq.2) u0=.25*u3+.5*u2+.25*u1
c
c       *** dump out sample by sample
c
        if(u0.gt.32767.) u0=32767.
        if(u0.lt.-32767.) u0=-32767.
        ix(1)=u0
c
```

```
c       *** output one speech sample from array ix(.)
c       *** subroutine spchout is not shown
c
220     call spchout(1,ix)
c
c       *** interpolation factor
c
        factor=ur*1/float(M)
        if(factor.gt.1.) factor=1.
c
c       *** rms interpolation
c
        rmsi=irms2+factor*(irms1-irms2)
c
c       *** pitch interpolation
c
        ipti=ipt2+factor*(ipt1-ipt2)
c
c       *** amplitude spectrum interpolation
c
        do 230 i=1,80
230     ampi(i)=amp2(i)+factor*(amp1(i)-amp2(i))
c
c       *** phase spectrum selection
c
        if(factor.gt..5) go to 235
        do 232 i=1,80
232     phasei(i)=phase2(i)
        go to 238
c
235     do 236 i=1,80
236     phasei(i)=phase1(i)
c
c       ----- inverse discrete Fourier transform -----
c
238     call idft(ipti,ampi,phasei,pw)
c
c       ----- if not pitch epoch -----
c
240     im=im+1
        xx(im)=pw(im)
300     continue
        go to 100
c
c       ------------------------------------------
c
999     end
```

```
c
c       ===== subroutines =====
c
c       ----- dc remove subroutine -----
c
        subroutine dcremove(a,b)
c
        b=(a-a1)+.9375*b1
        a1=a
        b1=b
        if(b.gt.32767.) b=32767.
        if(b.lt.-32767.) b=-32767.
        return
        end
c
c       ----- low-pass filter subroutine (-3 db at 1025 hz) -----
c
        subroutine lpf(r1,r2)
c
        y19=y18
        y18=y17
        y17=y16
        y16=y15
        y15=y14
        y14=y13
        y13=y12
        y12=y11
        y11=y10
        y10=y9
        y9=y8
        y8=y7
        y7=y6
        y6=y5
        y5=y4
        y4=y3
        y3=y2
        y2=y1
        y1=r1
        r2=.010*(y1+y19)+.013*(y2+y18)+.001*(y3+y17)-.024*(y4+y16)
&       -.045*(y5+y15)-.030*(y6+y14)+.039*(y7+y13)+.147*(y8+y12)
&       +.247*(y9+y11)+.285*y10
        if(r2.gt.32767.) r2=32767.
        if(r2.lt.-32767.) r2=-32767.
        return
        end
```

```
c
c       ----- pitch epoch finding subroutine -----
c
        subroutine centroid(i1,i2,ipp,i5lpf,small,loc)
        integer*2 i5lpf(1200)
c
        small=1000000.
        do 110 i=i1,i2
        sum=0.
        do 100 j=-ipp/2,-ipp/2+ipp-1
100     sum=sum+j*i5lpf(i+j)
        if(sum.gt.small) go to 110
        small=sum
        loc=i
110     continue
        return
        end
c
c       ----- pitch epoch refinement subroutine -----
c
        subroutine adjust(ipp,loc,i5lpf,small,sadj,locadj,ippadj)
        integer*2 i5lpf(1200)
c
        locadj=0
        Tprime=0
        sadj=1000000.
        irng=ipp/16
        do 110 i=loc-irng,loc+irng
        do 110 k=-irng,irng
        sum=0.
        do 100 j=-(ipp+k)/2,-(ipp+k)/2+(ipp+k)-1
100     sum=sum+j*i5lpf(i+j)
        if(sum.gt.sadj) go to 110
        sadj=sum
        locadj=i
        ippadj=ipp+k
110     continue
        return
        end
```

```
c
c       ----- discrete Fourier transform -----
c
        subroutine dft(ns,e1,amp,phase,nn)
        dimension e1(160),amp(80),phase(80)
c
        if(mod(ns,2).eq.0) nn=ns/2+1
        if(mod(ns,2).eq.1) nn=(ns+1)/2
        p=2.*3.1415926/ns
        tpi=2.*3.1415926
        tpit=tpi*(1./8000.)
        fs=8000./ns
c
100     do 110 j=1,nn
        rsum=0.
        xsum=0.
        const=tpit*fs*(j-1)
        do 120 i=1,ns
        arg=const*(i-1)
        rsum=rsum+e1(i)*cos(arg)
        xsum=xsum+e1(i)*sin(arg)
120     continue
        r=rsum/ns
        x=xsum/ns
        amp(j)=sqrt(r**2+x**2)
        phase(j)=atan2(x,r)
110     continue
        return
        end
```

```
c
c       ----- inverse discrete Fourier transform -----
c
        subroutine idft(ns,amp,phase,e2)
        dimension e2(160),amp(80),phase(80)
c
        if(mod(ns,2).eq.0) nn=ns/2+1
        if(mod(ns,2).eq.1) nn=(ns+1)/2
        p=2.*3.1415926/ns
        tpi=2.*3.1415926
        tpit=tpi*(1./8000.)
        fs=8000./ns
c
        amp(1)=.5*amp(1)
        if(mod(ns,2).eq.0) amp(nn)=.5*amp(nn)
        do 210 i=1,ns
        tsum=0.
        const=tpit*fs*(i-1)
        do 220 j=1,nn
        arg=const*(j-1)
        tsum=tsum+amp(j)*cos(arg-phase(j))
220     continue
        e2(i)=2*tsum
210     continue
300     return
        end
```

Navy Case No. 77,023

Pitch Period T 123

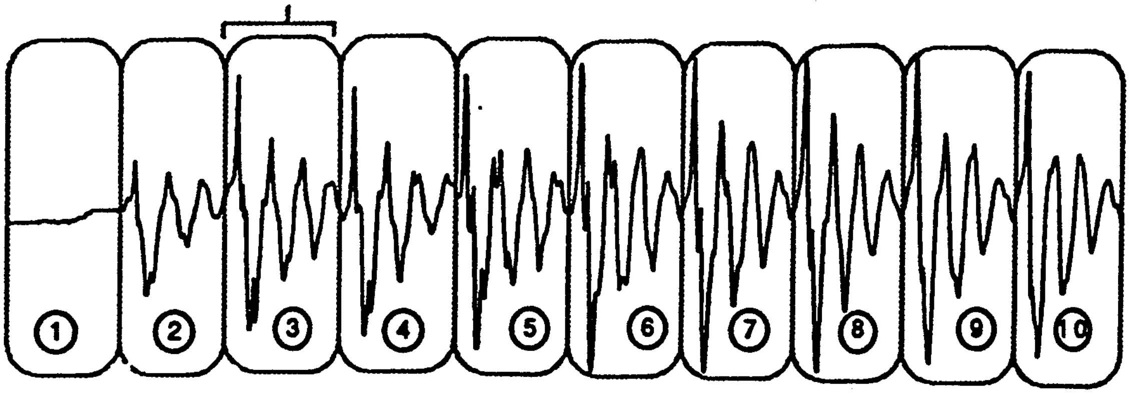
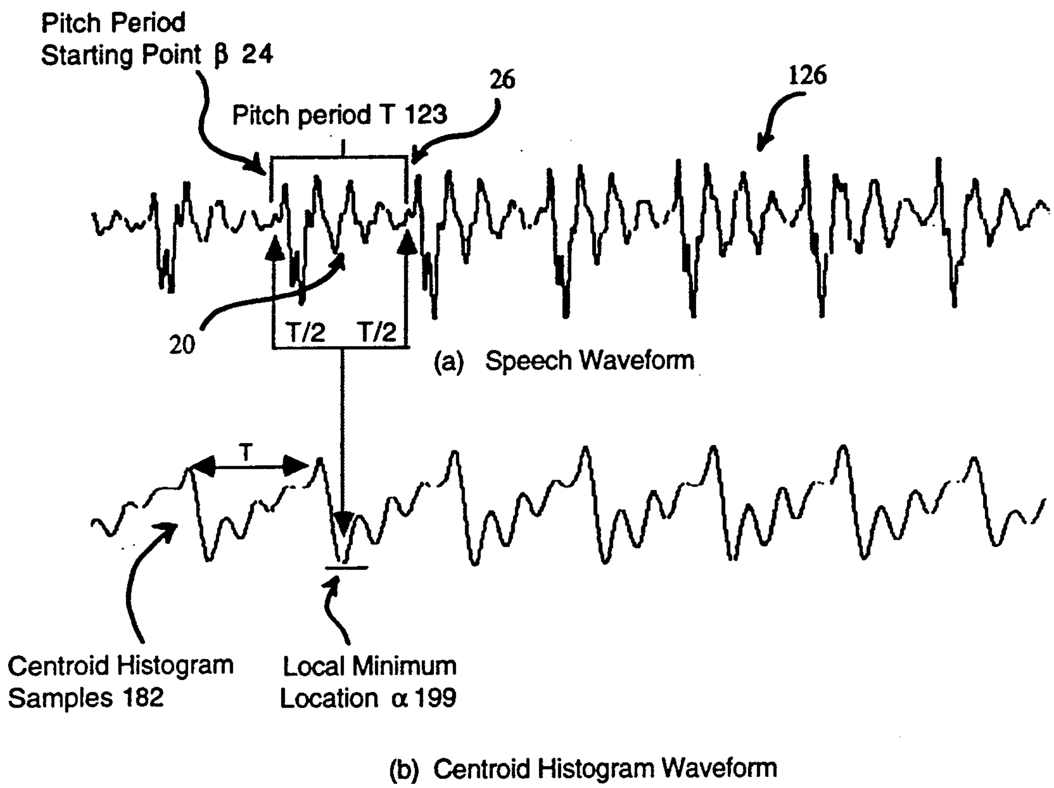


FIG. 1



Pitch Period
Starting Point β 24
26
126
Pitch period T 123
20
T/2
T/2
(a) Speech Waveform
T
Centroid Histogram
Samples 182
Local Minimum
Location α 199
(b) Centroid Histogram Waveform

FIG. 2

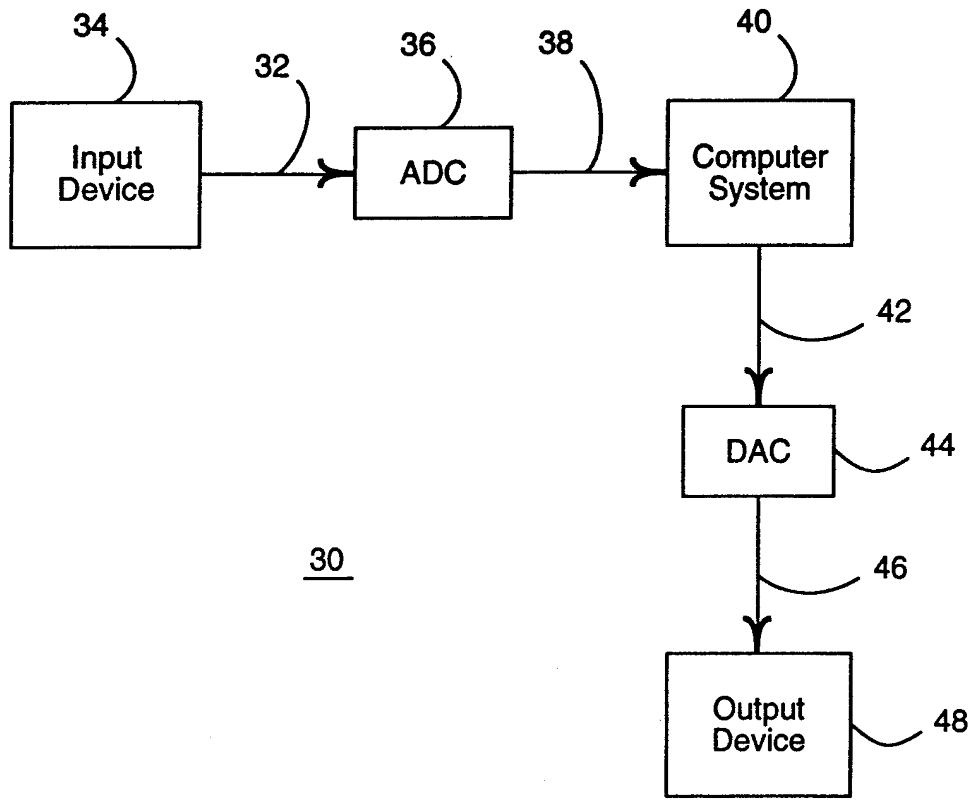
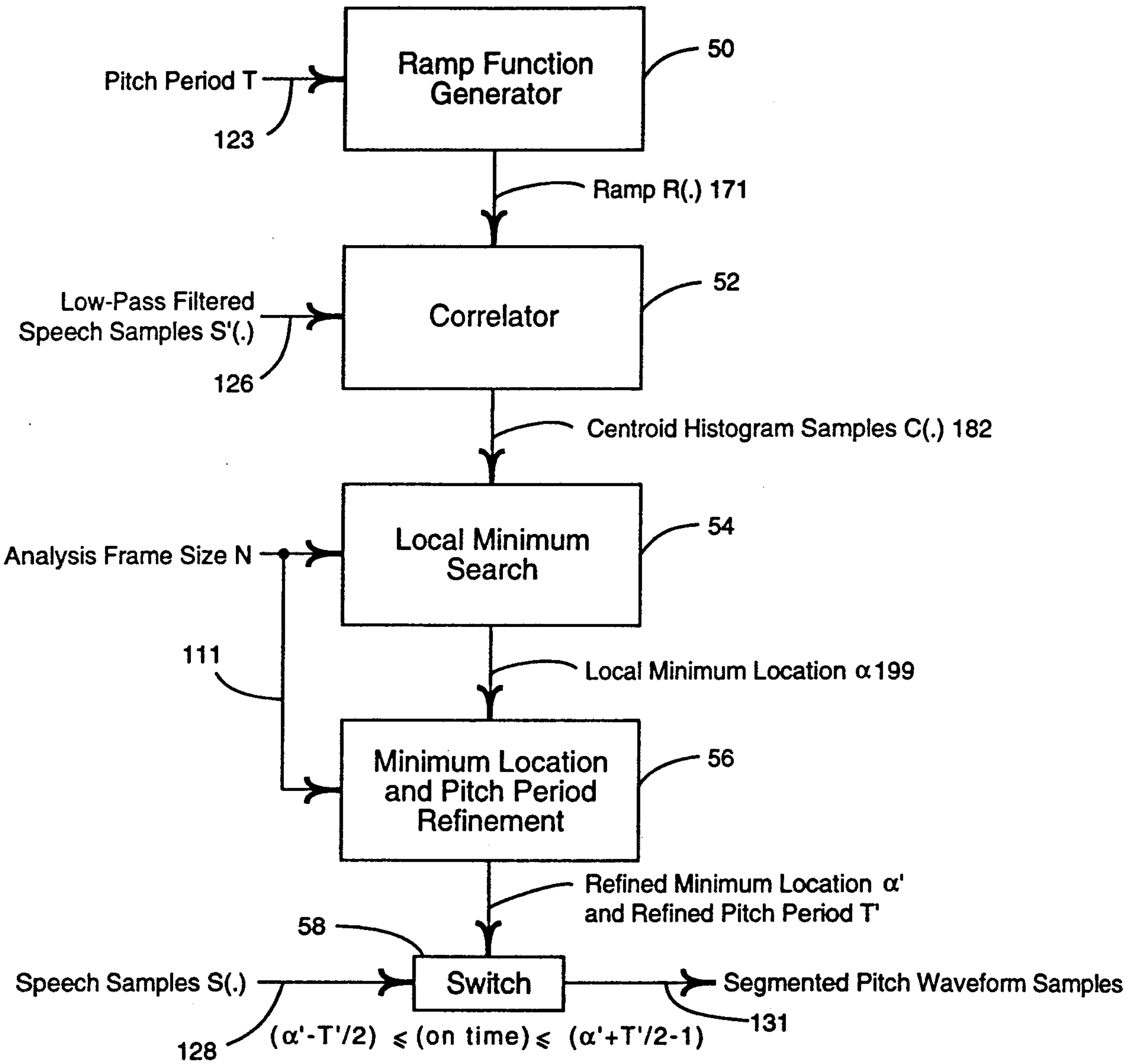
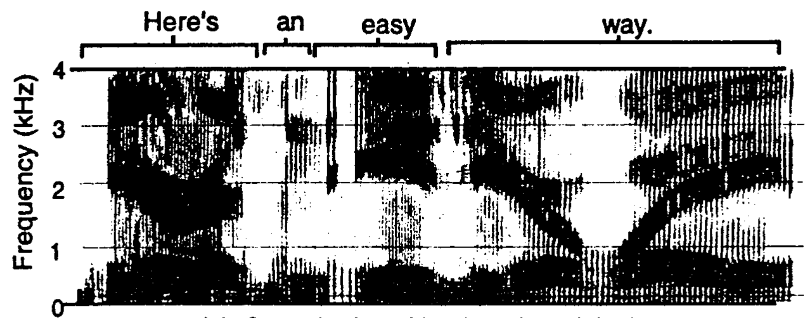

34
32
36
38
40
Input Device
ADC
Computer System
42
DAC
44
30
46
Output Device
48

FIG. 3



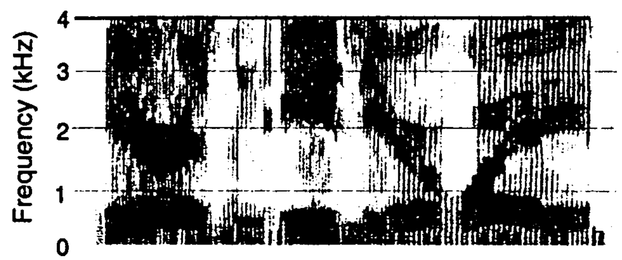
Pitch Period T
123
Ramp Function Generator
50
Ramp R(.) 171
Low-Pass Filtered Speech Samples S'(.)
126
Correlator
52
Centroid Histogram Samples C(.) 182
Analysis Frame Size N
Local Minimum Search
54
111
Local Minimum Location α 199
Minimum Location and Pitch Period Refinement
56
Refined Minimum Location α' and Refined Pitch Period T'
58
Speech Samples S(.)
Switch
Segmented Pitch Waveform Samples
131
128
(α'-T'/2) ≤ (on time) ≤ (α'+T'/2-1)

FIG. 4



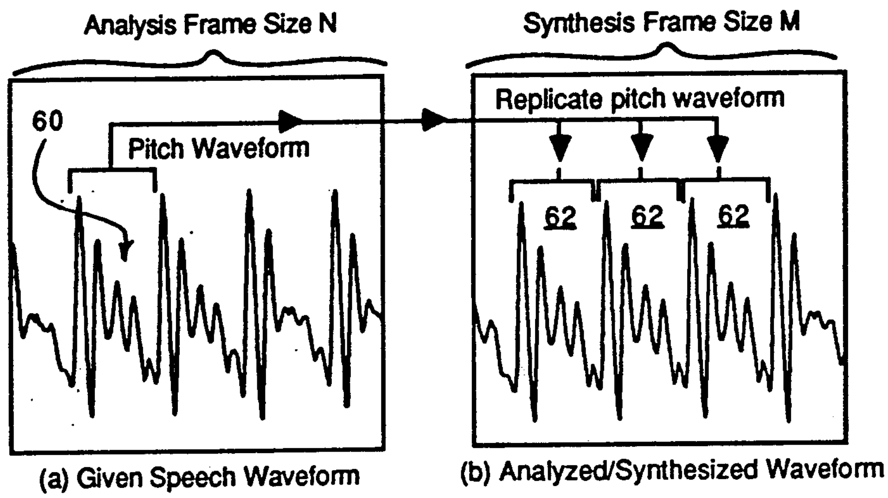
(a) Speech played back at the original rate

(b) Speech played back at 30% slower rate



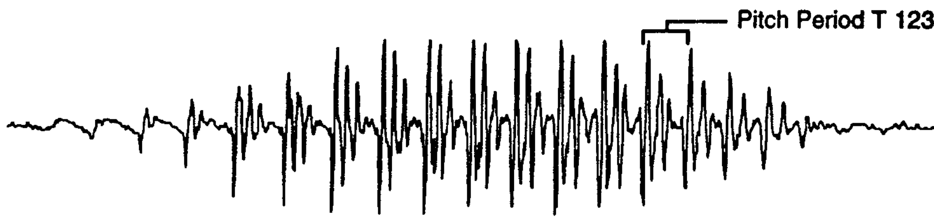
(c) Speech played back at 30% faster rate

FIG. 5

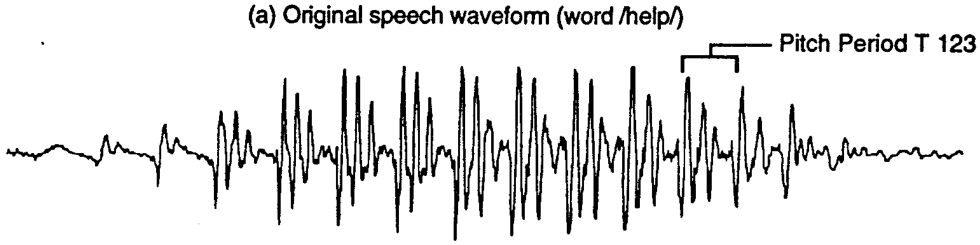


Analysis Frame Size N
Synthesis Frame Size M
60
Pitch Waveform
Replicate pitch waveform
62
62
62
(a) Given Speech Waveform
(b) Analyzed/Synthesized Waveform

FIG. 6



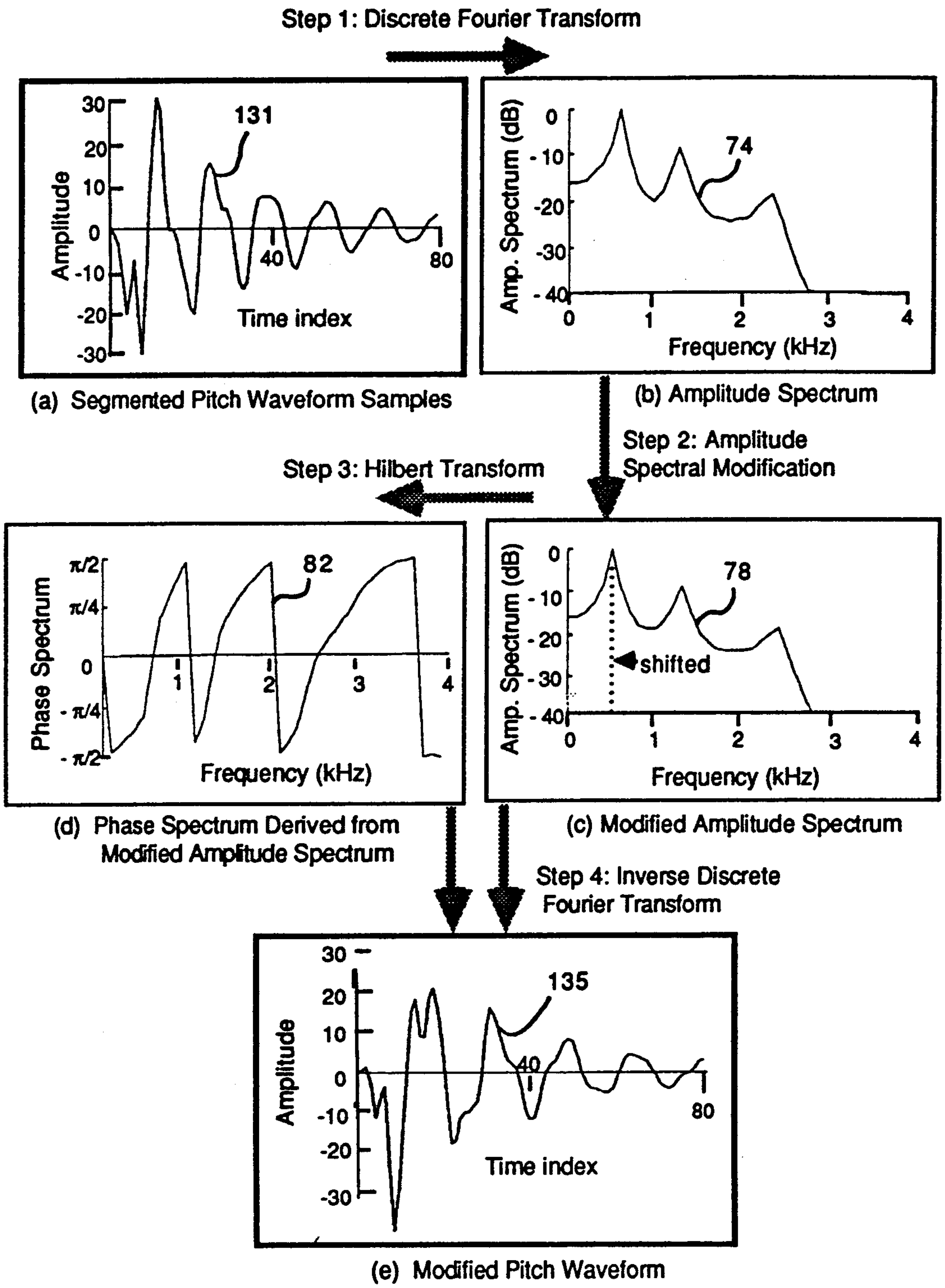
(a) Original speech waveform (word /help/)



(b) Original speech waveform with 30% lower pitch

(c) Original speech waveform with 30% higher pitch

FIG. 7



Step 1: Discrete Fourier Transform
30
20
10
0
-10
-20
-30
Amplitude
131
40
80
Time index
(a) Segmented Pitch Waveform Samples
0
- 10
-20
-30
- 40
Amp. Spectrum (dB)
74
0
1
2
3
4
Frequency (kHz)
(b) Amplitude Spectrum
Step 2: Amplitude Spectral Modification
Step 3: Hilbert Transform
π/2
π/4
0
- π/4
- π/2
Phase Spectrum
82
1
2
3
4
Frequency (kHz)
(d) Phase Spectrum Derived from Modified Amplitude Spectrum
0
- 10
- 20
- 30
- 40
Amp. Spectrum (dB)
78
shifted
0
1
2
3
4
Frequency (kHz)
(c) Modified Amplitude Spectrum
Step 4: Inverse Discrete Fourier Transform
30
20
10
0
-10
-20
-30
Amplitude
135
40
80
Time index
(e) Modified Pitch Waveform

FIG. 8

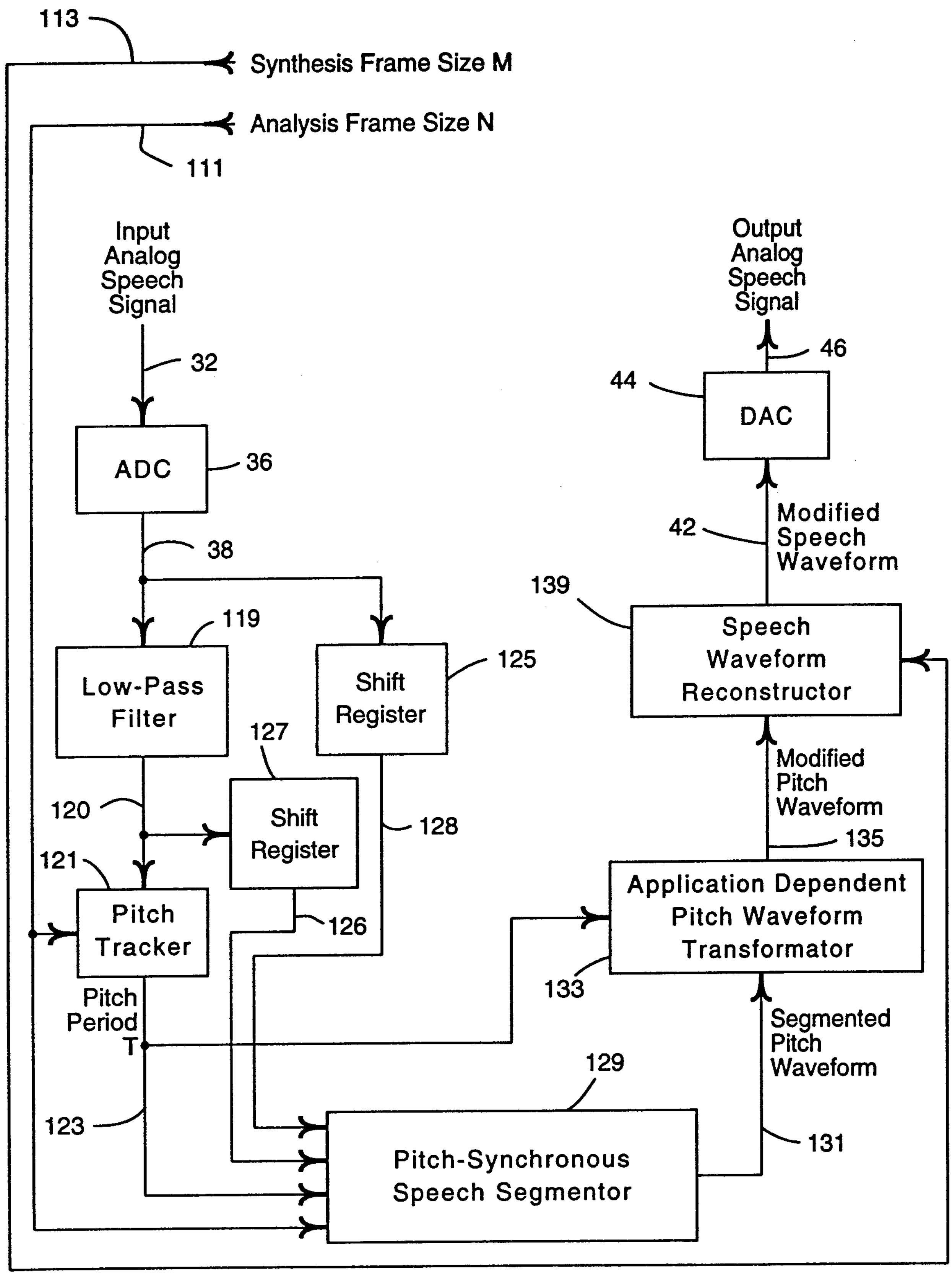
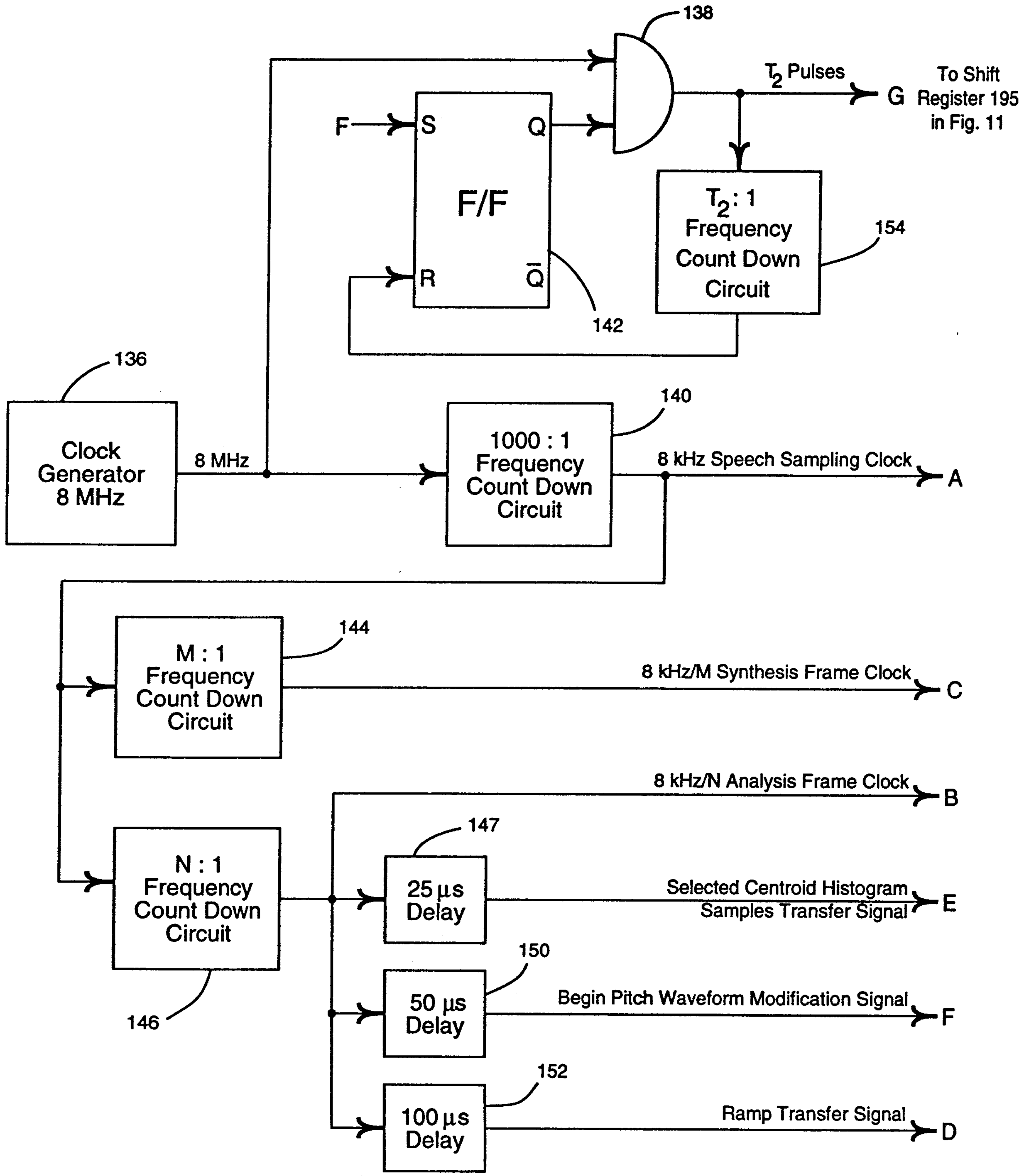
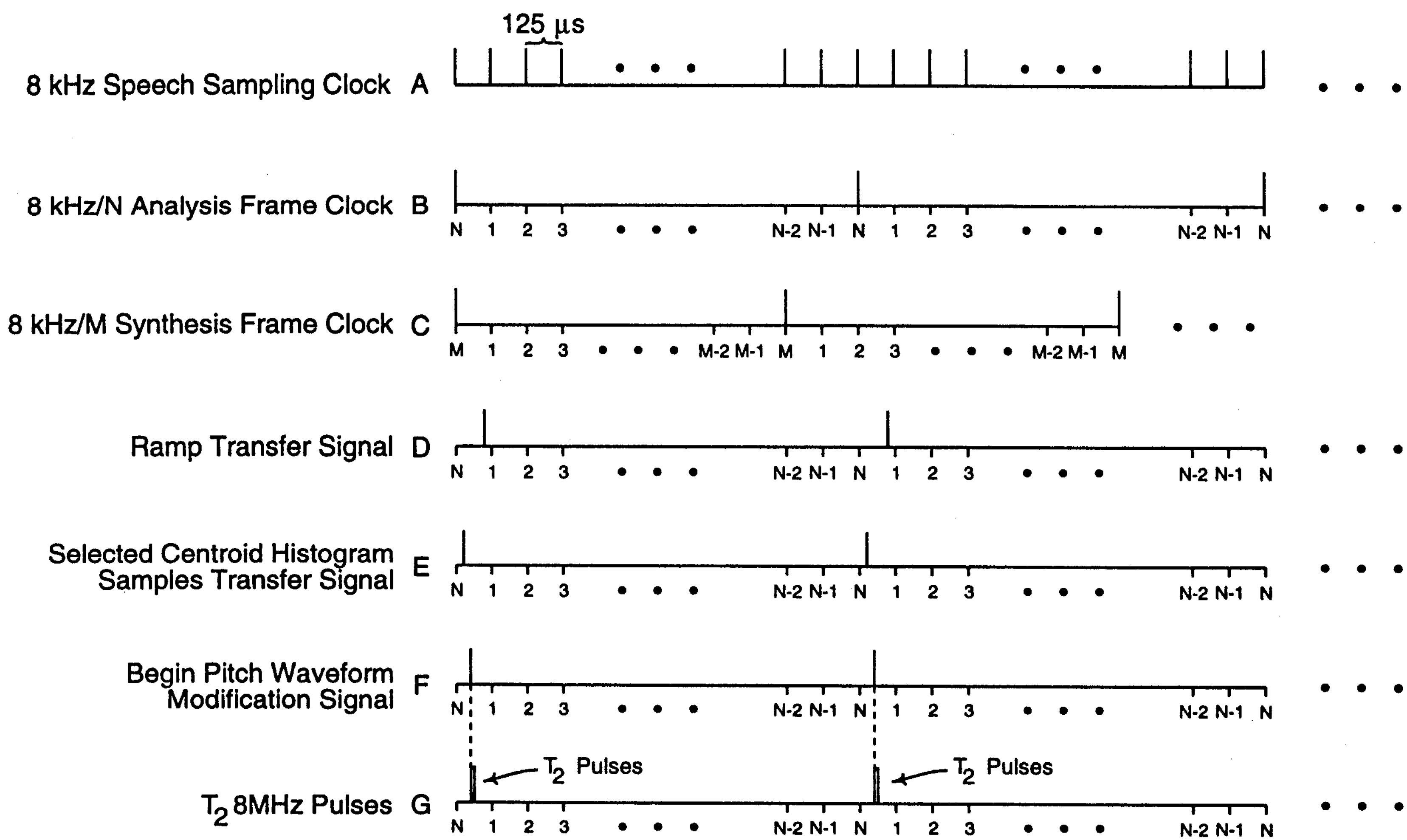

113
Synthesis Frame Size M
Analysis Frame Size N
111
Input
Analog
Speech
Signal
32
ADC
36
38
119
Low-Pass
Filter
Shift
Register
125
127
Shift
Register
120
121
128
Pitch
Tracker
126
Pitch
Period
T
123
Output
Analog
Speech
Signal
46
44
DAC
42
Modified
Speech
Waveform
139
Speech
Waveform
Reconstructor
Modified
Pitch
Waveform
135
Application Dependent
Pitch Waveform
Transformator
133
Segmented
Pitch
Waveform
129
131
Pitch-Synchronous
Speech Segmentor

FIG. 9



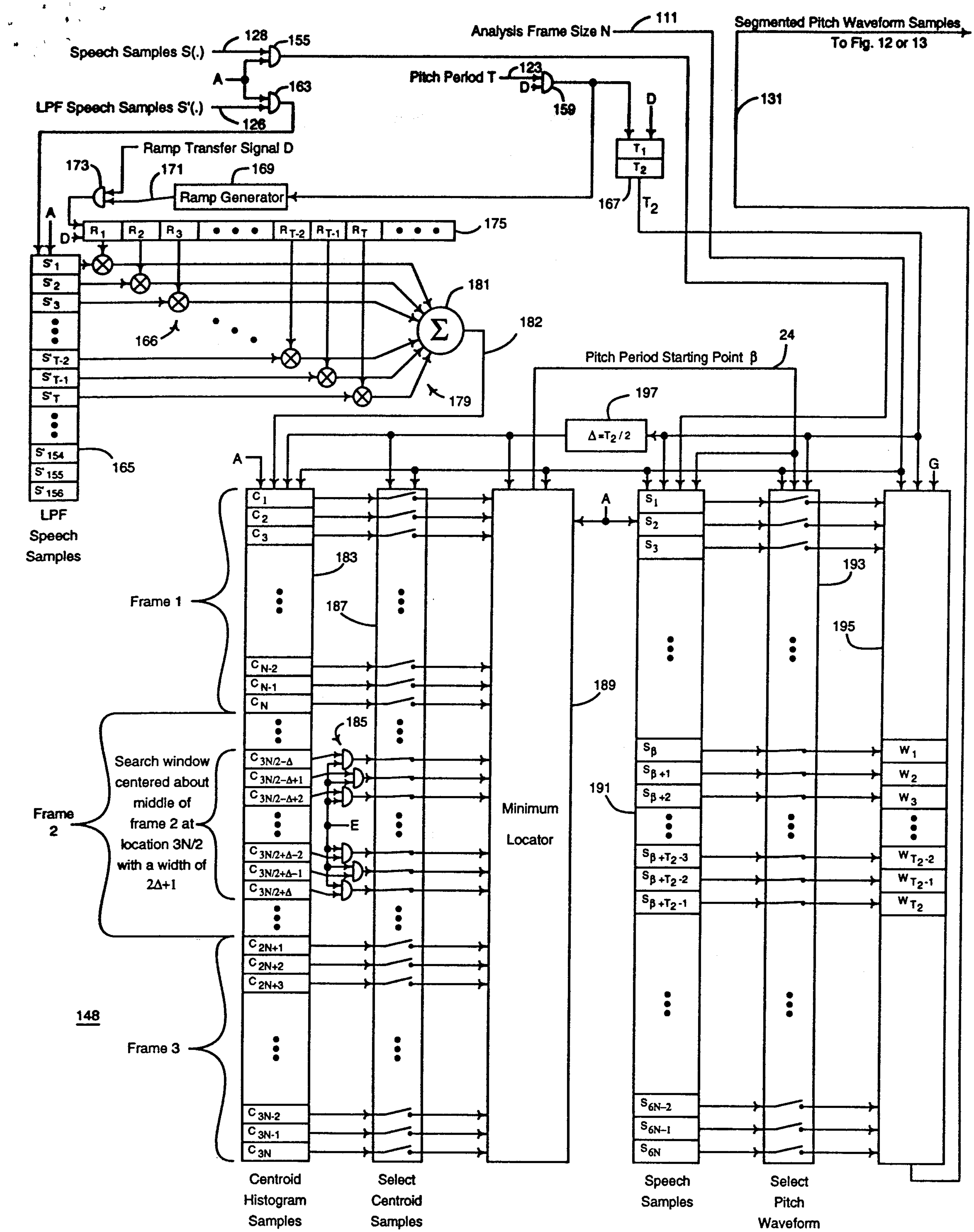
138
T_2 Pulses
G
To Shift Register 195 in Fig. 11
F
S
Q
F/F
R
$\bar{Q}$
142
T_2 : 1 Frequency Count Down Circuit
154
136
Clock Generator 8 MHz
8 MHz
140
1000 : 1 Frequency Count Down Circuit
8 kHz Speech Sampling Clock
A
144
M : 1 Frequency Count Down Circuit
8 kHz/M Synthesis Frame Clock
C
8 kHz/N Analysis Frame Clock
B
147
N : 1 Frequency Count Down Circuit
25 µs Delay
Selected Centroid Histogram Samples Transfer Signal
E
150
146
50 µs Delay
Begin Pitch Waveform Modification Signal
F
152
100 µs Delay
Ramp Transfer Signal
D

FIG. 10(a)



125 µs
8 kHz Speech Sampling Clock A
8 kHz/N Analysis Frame Clock B
8 kHz/M Synthesis Frame Clock C
Ramp Transfer Signal D
Selected Centroid Histogram Samples Transfer Signal E
Begin Pitch Waveform Modification Signal F
T2 8MHz Pulses G
T2 Pulses
T2 Pulses
N 1 2 3 N-2 N-1 N 1 2 3 N-2 N-1 N
M 1 2 3 M-2 M-1 M 1 2 3 M-2 M-1 M

FIG. 10(b)



Speech Samples S(.)
LPF Speech Samples S'(.)
Analysis Frame Size N
Segmented Pitch Waveform Samples To Fig. 12 or 13
Pitch Period T
Ramp Transfer Signal D
Ramp Generator
Pitch Period Starting Point β
Δ=T2/2
Frame 1
Frame 2
Frame 3
Search window centered about middle of frame 2 at location 3N/2 with a width of 2Δ+1
Minimum Locator
LPF Speech Samples
Centroid Histogram Samples
Select Centroid Samples
Speech Samples
Select Pitch Waveform
148

FIG. 11

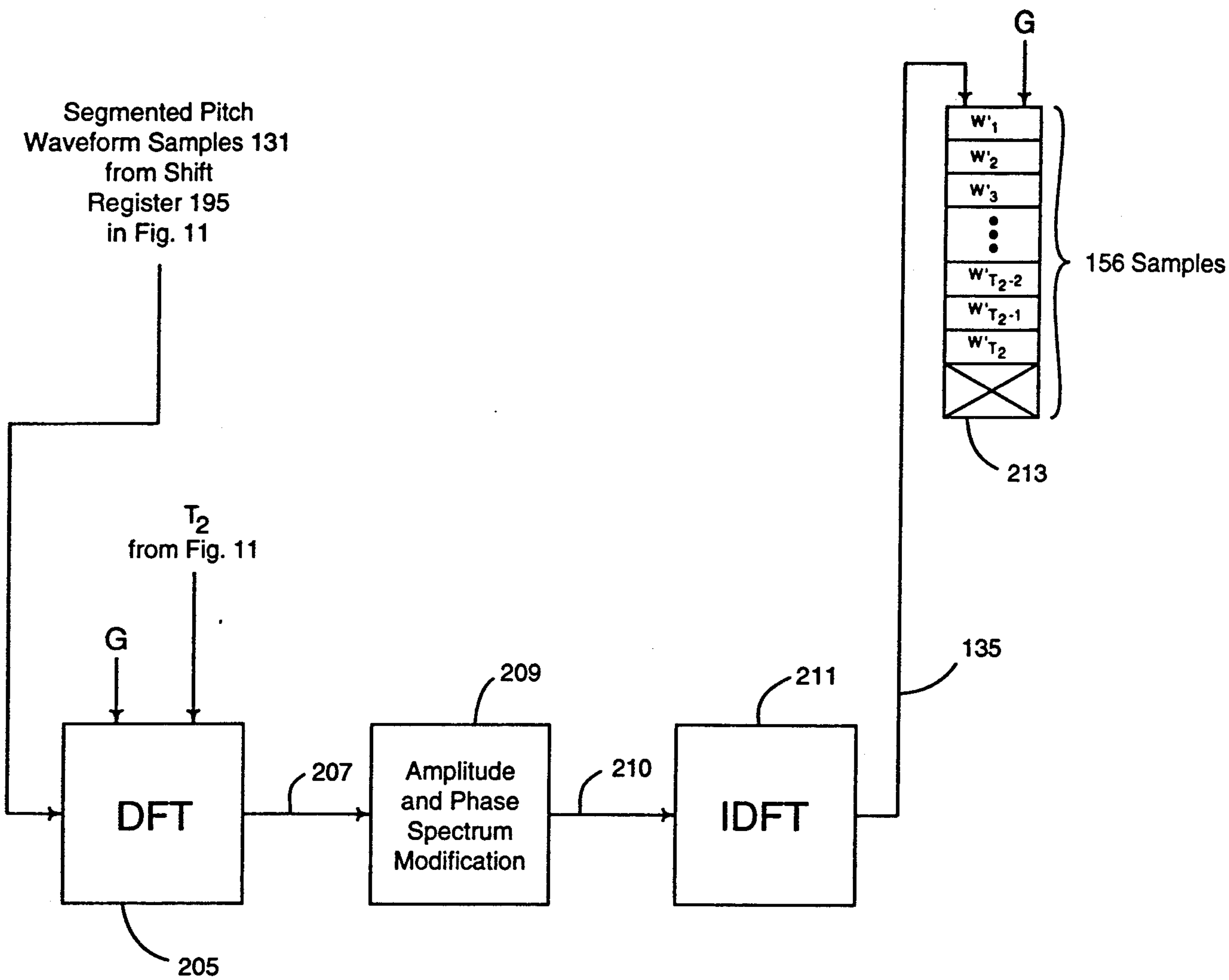
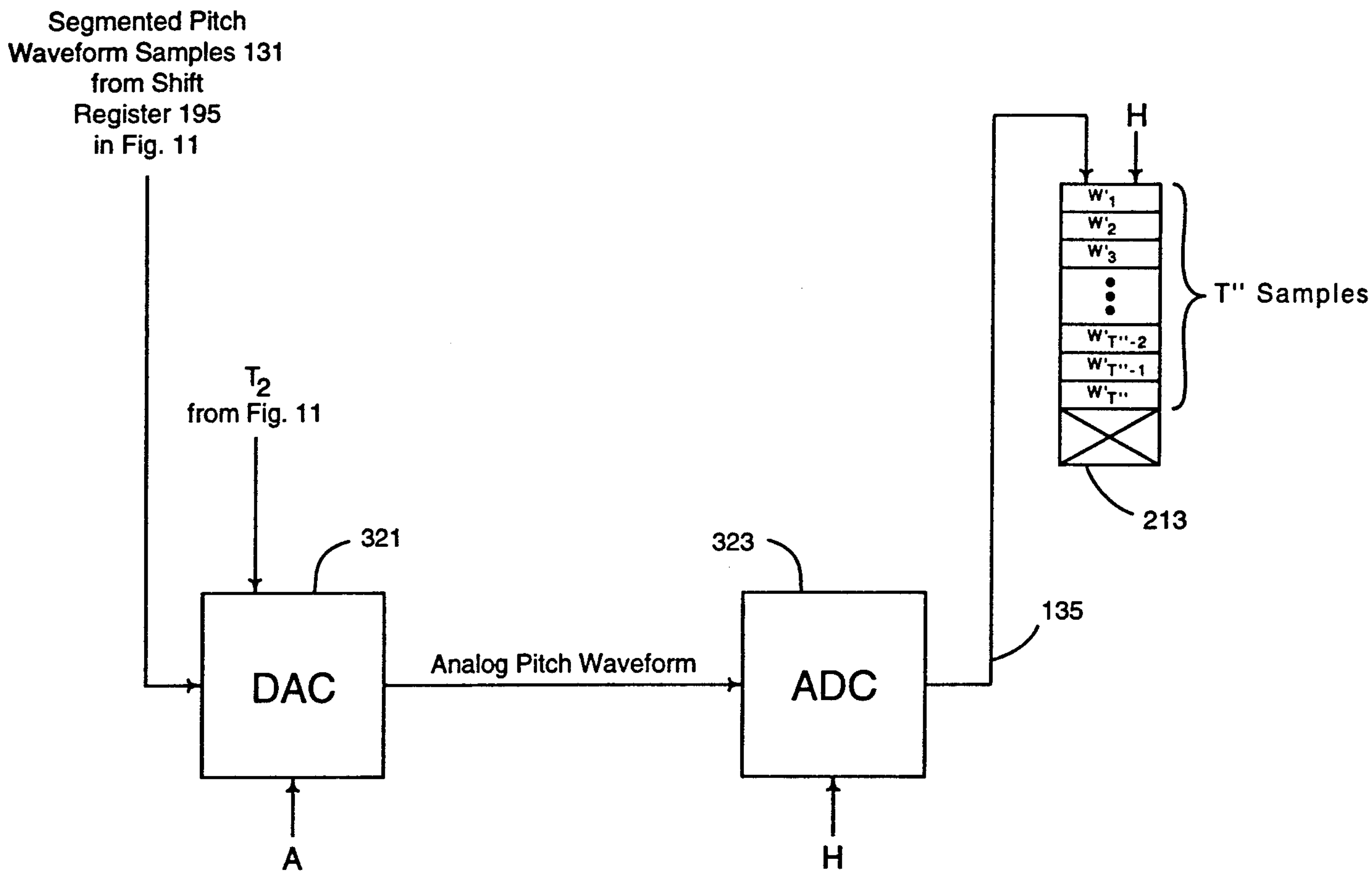

Segmented Pitch Waveform Samples 131 from Shift Register 195 in Fig. 11
G
T_2 from Fig. 11
G
DFT
205
207
Amplitude and Phase Spectrum Modification
209
210
IDFT
211
135
W'_1
W'_2
W'_3
W'_{T_2-2}
W'_{T_2-1}
W'_{T_2}
156 Samples
213

FIG. 12



Segmented Pitch
Waveform Samples 131
from Shift
Register 195
in Fig. 11
T2
from Fig. 11
321
DAC
A
Analog Pitch Waveform
323
ADC
H
135
H
W'1
W'2
W'3
W'T''-2
W'T''-1
W'T''
T'' Samples
213

FIG. 13

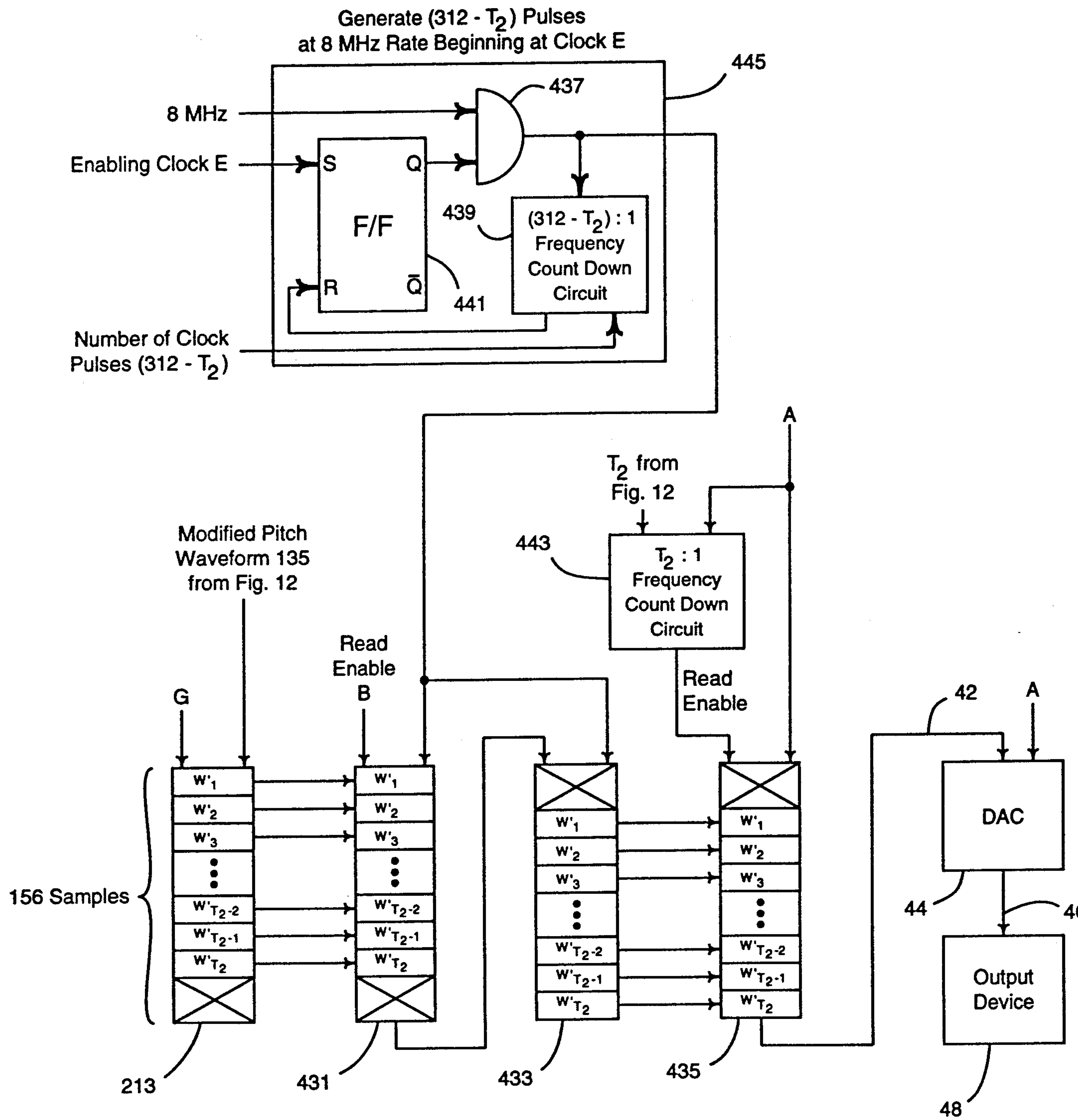

Generate (312 - T2) Pulses at 8 MHz Rate Beginning at Clock E
445
437
8 MHz
Enabling Clock E
S
Q
F/F
R
Q̄
439
441
(312 - T2) : 1 Frequency Count Down Circuit
Number of Clock Pulses (312 - T2)
A
T2 from Fig. 12
443
T2 : 1 Frequency Count Down Circuit
Modified Pitch Waveform 135 from Fig. 12
Read Enable B
G
Read Enable
42
A
DAC
44
46
Output Device
156 Samples
W'1
W'2
W'3
W'T2-2
W'T2-1
W'T2
213
431
433
435
48

FIG. 14

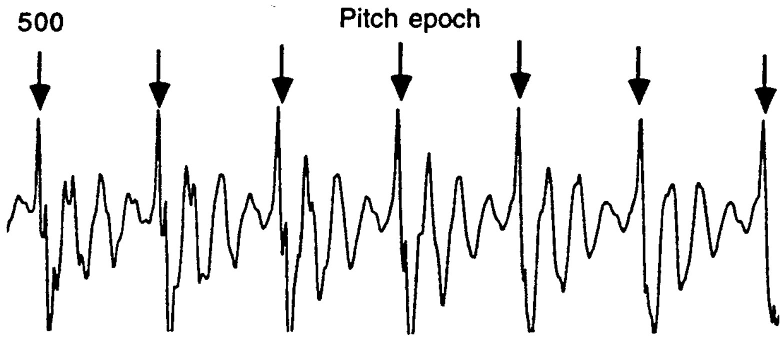
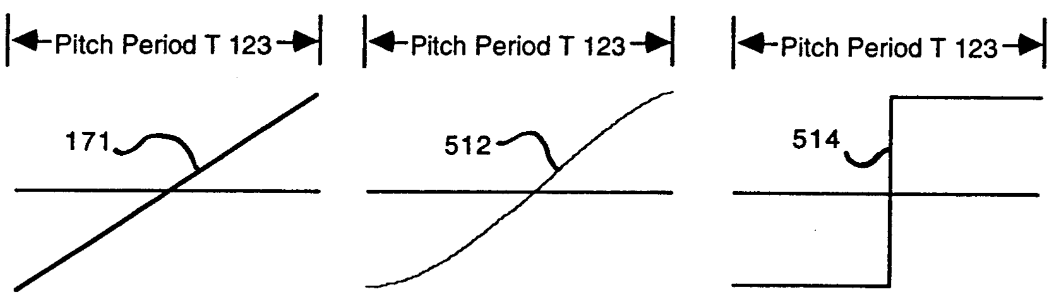


FIG. 15



(a) Ramp R(.)

(b) Sine Wave from -π to π (Alternative)

(c) Step Function (Alternative)

FIG. 16